Visualisation de données multivariées : réimplémentation des fonctionnalités graphiques de la librairie `ade4`

A. Julien-Laferriere[a] and **S. Dray**[b]

[a]Laboratoire de biométrie et biologie évolutive (UMR CNRS 5558)
CNRS - Université Lyon 1
43 bd du 11 novembre 1918, 69622 Villeurbanne, France
alice.julien-laferriere@univ-lyon1.fr

[b]Laboratoire de biométrie et biologie évolutive (UMR CNRS 5558)
CNRS - Université Lyon 1
43 bd du 11 novembre 1918, 69622 Villeurbanne, France
stephane.dray@univ-lyon1.fr

Mots clefs : Analyse multivariée, Graphique, Visualisation.

Dans un grand nombre de disciplines (e.g., écologie, génétique, santé), les récents développements technologiques facilitent la collecte et la gestion de données et conduisent à l'élaboration de bases de données massives dont la structure est de plus en plus complexe (multivariée, hiérarchisée, structurée dans l'espace et/ou le temps, etc.). L'analyse et la représentation de ces données nécessitent des méthodes adaptées prenant en compte leurs caractéristiques intrinsèques. Dans ce contexte, les méthodes d'analyse multivariée fournissent un ensemble d'outils permettant de résumer l'information contenue dans de grands tableaux en identifiant les relations entre variables, et les similarités entre individus. Les résultats sont alors présentés sous la forme de graphiques, pour un nombre réduit de dimensions, permettant une exploration des principales structures identifiées dans les données.

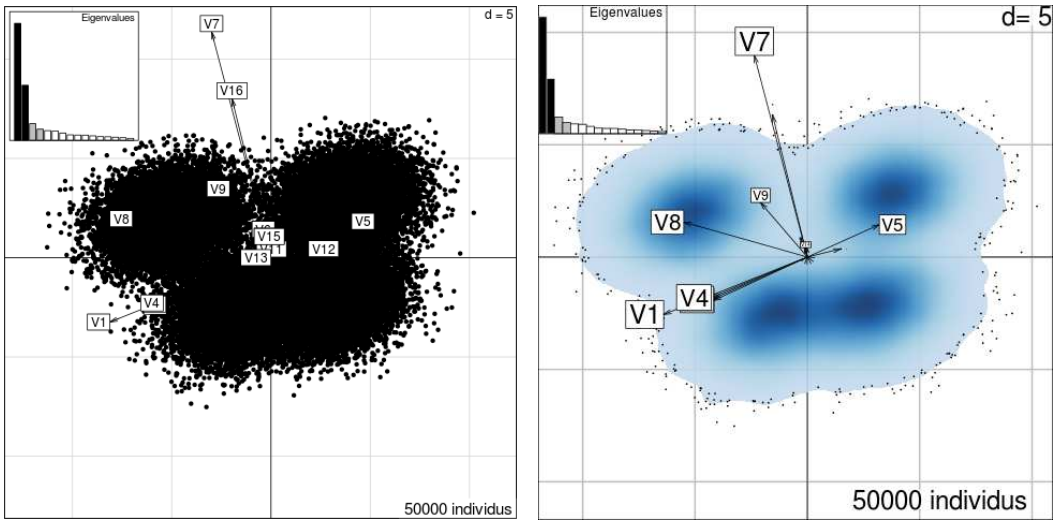
Depuis 2002, le package R `ade4` [1], développé au laboratoire de Biométrie et Biologie Évolutive, fournit un ensemble de méthodes permettant l'analyse d'un seul, de deux mais aussi de K tableaux. A ce jour, une quarantaine de méthodes différentes ont été implémentées, dont près de la moitié ont été développées par les auteurs du package. Une quarantaine de fonctions graphiques sont également disponibles afin de représenter les résultats issus de ces analyses. Près de dix ans après la première distribution d'`ade4` sur les serveurs du CRAN, nous sommes en train de mettre en place de nouvelles modalités de représentation graphique permettant une utilisation plus souple et plus intuitive du logiciel. L'objectif est d'améliorer la visualisation des données et/ou des résultats d'analyses en s'appuyant sur les nouvelles fonctionnalités offertes par R.

Cette nouvelle implémentation est réalisée avec une programmation orientée objet (S4) en s'appuyant sur une hiérarchisation des différentes représentations graphiques disponibles. Les graphiques sont alors stockées sous la forme d'objets et il est ainsi possible de les créer sans les visualiser ou de les manipuler *a posteriori*. Ces objets peuvent être combinés (juxtaposition, superposition) afin d'observer, dans une même fenêtre graphique, différents niveaux d'information.

Ces nouvelles fonctionnalités graphiques s'appuient sur l'utilisation du package `lattice` [2] qui permet d'obtenir une grande souplesse dans la production de graphiques conditionnels pour

l'exploration de données multi-dimensionnelles.

En tirant profit des fonctionnalités implémentées dans `lattice`, nous avons mis en place deux grandes classes d'objets spécifiquement associées à la représentation graphique de données sous `ade4`. La première nous permet de définir une série de graphiques élémentaires utilisés en analyse multivariée alors que la seconde classe permet de gérer une collection de graphiques obtenus par superposition ou juxtaposition. De plus, de nombreux paramètres sont disponibles afin de personnaliser facilement les graphiques et mettre ainsi en relief les principales structures associées à un jeu de données particulier (partition en groupe, structure spatiale, etc.).



(a) (b)

FIGURE 1 – Exemple et comparaison de graphiques obtenus dans `ade4` : projection des variables et individus sur les deux premiers axes d'une analyse en composantes principales (ACP). (a) Ancienne implémentation, (b) Nouvelle implémentation. En haut à gauche sont représentés les valeurs propres de l'analyse.

Sur la figure 1 sont représentés une partie des résultats d'une analyse en composantes principales sur des données fictives comportant seize variables et cinquante mille individus. La figure obtenue avec l'ancienne implémentation (Fig. 1a) ne permet pas d'observer clairement la distribution des individus. Sur la figure 1b, un des nouveaux graphiques disponibles permet de représenter les individus non par des points mais par une nappe de densité qui permet de mieux rendre compte de la distribution de ceux-ci. Enfin, ici, le graphique a été personnalisé pour augmenter la taille du titre mais aussi changer la taille des étiquettes des variables qui est proportionnelle à leurs contributions sur les deux axes représentés.

Ce travail constitue donc une étape importante pour le package `ade4` en améliorant sensiblement les fonctionnalités actuelles et en offrant un cadre général et flexible facilitant l'implémentation de futurs outils graphiques.

Références

[1] S. Dray, and A.B Dufour. The ade4 package : implementing the duality diagram for ecologists *Journal of Statistical Software*, 22(4) :1–20,2007

[2] D. Sarkar. Lattice : multivariate data visualization with R *Springer Verlag*, 2008